DOCUMENT RESUME

ED 093 892 SP 008 263

AUTHOR Rosenshine, Barak
TITLE Issues and Research Suggestions in Classroom Observation.
PUB DATE Nov 73
NOTE 10p.; Paper prepared for the Conference on Observational Techniques, Early Learning Task Force, National Institute of Education (Washington, D.C., November 1973)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Classroom Observation Techniques; *Classroom Research; Classrooms; Codification; Educational Research; Indexing; Measurement; Performance Criteria; Teacher Evaluation

ABSTRACT

This paper examines current forms and instruments of classroom observation and suggests directions for future research studies. The following topics are covered: the need for good criterion measures, the use of observational instruments in outcome measures, the development of procedures to code the content that is covered in a classroom, the dangers of excessive complexity, the use of observation for teacher competency assessment, naturalistic observation, the coding of questions and cognitive interactions, a typology of questions and cognitive interactions, the indexing of implementation, and a proposal for a data bank for secondary analyses. (JA)

Issues and Research Suggestions in Classroom Observation[1]

Barak Rosenshine
University of Illinois

1. The need for good criterion measures.

Many of the questions in the use of observation instruments are answerable if one can test the functional relationships between events recorded through these instruments and student outcomes. Unfortunately, acceptable measures do not exist for many important educational outcomes; the testing of functional relationships is limited by the current inadequacy of the outcome measures. Thus, there is a need for research, development, and reviews on outcome measures.

2. Can observational instruments be used as outcome measures?

Yes, but they need to be used in situations that are relatively independent of the classroom, otherwise they would be measuring process and not outcomes. Observational instruments have been used to assess criterion behavior in correlational and experimental studies. For example, the Russell Sage Social Relations Test employs coding of pupil interaction as a group of students work on construction-block projects. Experimental research in early childhood has employed using observational data of the child in his environment as both baseline and posttest data while the treatment took place in a special setting.

But using process measures taken in class as outcome measures for variables such as independence, curiousity, cooperation or persistency may lead to unwarranted conclusions. Currently, although there have been a number of studies relating process measured to student gain in reading, the correlations have not been particularly high despite the fact that investigators chose variables which they expected would be strongly related to reading gain. If, as yet, we are unable to establish strong relationships in process-outcome studies, then it does not seem legitimate to claim that certain processes --- which have not yet been related to outcomes -- are important in their own right.

---

[1] Paper prepared for Conference on Observational Techniques, Early Learning Task Force, National Institute of Education, Washington, D.C., November, 1973.

There are probably a few process variables which are important in their own right. Teacher behaviors which demean or humiliate pupils are currently undesirable no matter what relationship is demonstrated between process and outcome. (Even in such a case one would want to know if the pupils felt humiliated; what do we do if the observer believed that the behaviors humiliated pupils but the pupils did not feel the same way?). There are other teacher behaviors which people appear to advocate on grounds of taste. Arguments about the extent and type of individualization, the choice and method of studying various subject areas, and the necessary amount of joy in a classroom appear based on grounds of taste rather than moral grounds or research grounds. I assert that just as school dress codes cannot be justified by taste alone, teachers cannot be held accountable for specific classroom transactions solely on grounds of taste.

In the introduction to his book on testing, Ebel presents another illustration of the process-outcome measurement issue. He writes that judges watching children at play could make estimates of the relative abilities of the children to run fast, jump high, or throw an object far. He argues that everyone concerned would probably prefer to see these estimates made under some standardized and controlled, if somewhat artificial, conditions of a regular track meet.

3. There are many forms of observation.

It would be a mistake to limit observational instruments to category instruments. At present, rating instruments, teacher self reports, sign instruments, and student questionnaires are all viable observational instruments. At present we do not know whether one form is more functional than another and this is probably a poor question. That is, some forms may be more functional for some constructs (e.g. teacher positive responses) and other forms more functional for constructs such as type of question or interaction.

4. One does not validate an observational instrument.

Even in research which looks at functional relationships, one can only begin to validate items on an observational instrument, not the complete instrument.

5. The coding of content covered.

The development of procedures to code the content which is covered in a classroom is a research need of the highest priority. At present, there are only three or four observation instruments which include codes on content. In almost all current observational instruments teacher divergent questions on how to arrange a classroom, for example, receive identical coding as questions on the application of a principle to new situations.

The RAMOS instrument, developed by R. Calfee and K. Hoover, is one example of a new instrument with a content dimension. The reading dimension

contains sixteen options such as simple decoding, syllabification, recreational reading, comprehension of relations, and comprehension of sequence. This dimension could be used with any category or sign instrument so that the context of the behavior could be coded with the behavior. Some hypotheses of critical interest would be:

a) the correlations between teacher behaviors (or student behaviors) and the outcome measure(s) will be strengthened if the behaviors and the content area are coded together;

b) frequency counts of content behaviors alone will yield a substantial correlation with pupil outcome measure(s).

6. The danger of The Great Complexifiers.

In research on functional relationships it is easy to pose so many questions and issues that a researcher and a research enterprise can become immobilized. The Great Complexifiers are those who pose these additional questions, much as professors at a preliminary oral keep asking "Have you controlled for....".

Suppose one wished to mount a series of studies to look at teacher questions and student achievement in grades K-3. One could set up a matrix in which one factor consisted of the four grade levels, and a second factor was on the subject areas: reading, math, science, social science. Thus one begins with a fairly complex sixteen cell matrix.

The great complexifiers respond that the number of factors and cells are too few. They suggest a location factor, suggesting that the schools be divided into urban, suburban, and rural. They suggest an income factor, dividing pupils into low income and middle income. They suggest separate cells for male and female pupils; they suggest that race and ethnic background be considered, so that pupils are classified as black, Mexican, Puerto Rican, American Indian, Appalachian, and white. The complexifiers further suggest that tests are not unidemisional, and therefore outcomes should at least be classified as recall and processing outcomes. Finally, another complexifier will claim that there is no such thing as "first grade reading," but rather, there is MacMillan reading, Sullivan reading, Bank Street reading, and five other kinds of reading curriculum.

So to a sixteen cell matrix one adds three levels of school location, two income levels, two sex levels, six ethnic levels, two outcome levels, and eight curriculum levels for a 4 X 4 X 3 X 2 X 2 X 6 X 2 X 8 matrix of 18,000 cells!

Complexification and "did you control for" is just plausible enough to serve to immobilize a researcher or a research program because many of the variables or possible interest are not being studied (or can't be studied given the number of cells compared to the possible number of classes).

There is thus an urgent need to unsnarl the complexifiers and reach some consensus on which of all possible cells appear more promising than others. Again, the absence of accepted criteria measures will hamper such a venture, but I recommend that an effort be begun to sort out which pieces of this complexity seem worthwhile for future research.

7. Can observation be used for teacher competency assessment or for pupil competency assessment?

The use of observation for generic teacher competency assessment is unfeasible at this time because we know so little of functional relationships between behaviors and outcomes. A similar argument appears to apply for pupil assessment.

But assessment is possible within the content of curriculum or program implementation. In this case the criteria for assessment variables is one or more steps removed from outcome measures; in this case the criteria are defined by the developer and represent those actions considered important to implement the program according to the intentions of the developer(s).

Assessment of curriculum implementation, at this point, is not assessment of teacher competence or even of program competence. Rather, it is a necessary first step in planning research. Without subsequent research implementation assessment is not particularly meaningful because implementation variables are only hypotheses that these variables are important for the outcomes.

8. The importance of naturalistic observation.

Naturalistic observation can serve as both a source and as a supplement for categorical observation. Naturalistic observation can serve to suggest potentially functional variables which might have been overlooked when developing a category instrument.

The danger in both naturalistic and categorical observation alone is that an observer, researcher, or reader is too easily persuaded that the variables which strike him as important are indeed functional.

9. Developing "clean" observational concepts.

There are many problems which militate against developing clean observational concepts. The first is that there is too much noise to permit clean translation of concepts developed outside a classroom into an observational instrument. For example, consider a construct such as an "analysis question" taken from the Bloom et al Taxonomy, or a "divergent question" taken from Guilford's research. These constructs were operationalized in written questions. In all probability, these constructs do not fit neatly when coding actual questions in the classroom because there is too much noise.

For example, many classroom questions are attempts to get a pupil to clarify his response to the first question, or attempts to probe into an initial answer. Yes, the concepts developed by Bloom or Guilford do not fit a "probing" situation because their concepts were not developed for interactive settings. At present, despite the relative purity of the origins of the typologies developed by Bloom or by Guilford we have a great deal of trouble translating ideas developed from one source into the classroom. Another problem with Bloom's or Guilford's typology of questions is that there are so many existing typologies. In addition to these two, questions have been classified into six or more types by B.O. Smith, by Taba, by Brophy, and by Gallagher. We don't know how these different typologies converge, how they differ, and which categories of questions are functional.

10. Coding questions and cognitive interactions.

The following suggestions for research studies may further illustrate the difficulty of obtaining clean observational concepts.

Suppose one wished to develop a series of studies on the functional value of different ways of coding questions (or cognitive interchanges). As noted above, questions have been coded into six or more different types by a number of investigators such as Smith, Brophy, Bellack, Bloom, Taba, Gallagher, and Conners and Eisenberg. At present we do not know whether these concepts are similar or different, nor do we know the functional value of these concepts.

One way of developing research studies in this area would be to take three or four sets of specimin tapes and code them using the different ways of coding questions. Bob Soar alone may have enough sets of specimin tapes for one such study. Soar has audio tapes and a number of outcome measures for over 100 K-3 classrooms in Follow Through for at least two years.

Assuming that Soar's audiotapes represent specimin sets, one could code the tapes using each of the above seven categorizations and relate the obtained frequencies to measures of student achievement. The results would not validate a particular coding procedure, but they might tell us which specific items were more functional than others in this context. Studies of the intercorrelations among question types within coding schemes and across coding schemes could indicate how the question types cluster into independent groupings. The results obtained on one set of tapes could be cross validated against another set.

Whether this approach would yield conceptual clarity and stable functional relationships is testable. An alternative hypothesis would be that there are so many ways of developing a coding scheme based on each of the above categories of question types that the number of studies which could be run (and the number of valid and spurious correlations which could be obtained) makes this approach unmanageable.

Within any one set of question codes one still has questions on

coding single events or coding sequences
the unit of analysis (e.g. frequency, move, utterance, cycle, topic, etc.)

the number of different questions which fit into one question type
the number of dimensions (e.g. speaker, tone, content) to include within a count
the scale to be used to estimate frequency (e.g. category or sign method)

Whether different procedures for using the same concepts of questions will yield different, consistent, and functional results is a research question. Although I would guess that the results will be uninterpretable, I would recommend that this series of studies be run in the hopes of determining whether there are empirical procedures which might yield conceptual clarity.

The list above of issues in the technology of coding questions are independent of the theoretical origins of a set of categories of question types. Even if one decides to take the variables and their names from the theory and research of Dewey, Piaget, Miller, or Skinner, one still has to make decisions on the unit of analysis, the number of type of coding dimensions, and other issues. There are no guidelines for these decisions no matter how clean the theoretical origin of the observational instrument.

10a. A typology of questions and cognitive interactions.

As much as the idea of a typology of questions has appeal, the development of such a typology is difficult to conceive because of the variety of ways which exist to code questions (e.g. the codings developed by such as Smith, Bloom, Taba, and others) and the variety of recording procedures which might be used. The use of a data bank of interactions and outcomes to test which questions types and which recording procedures are most useful seems appealing, but I worry that the results of such a series of studies will not yield clean results. My worry, however, is testable.

10b. Determining functional units, approaches, and recording procedures.

If one returns to the example of recording questions or cognitive interactions, there remain a number of unresearched research issues. When developing an observational instrument, one must make decisions on:

the number of different behaviors to be included in a variable (e.g. are all instances of praise to be considered as one variable, or will subdivisions be made for different apparent forms of praise; similar for criticism, feedback, types of questions)

the number of dimensions to be encoded with each behavior (such dimensions could include the content, the source, the number of students attending, the firmness of the interaction, the role the teacher was in, additional cognitive and affective dimensions to an interaction)

the unit of recording (e.g. natural unit, simple count, sign count, rating)

how many sequences should be recorded (e.g. single instances, diads, triads)

whether smaller variables should be combined for analyses

whether ratios of behaviors should be used for analyses

The above list seems awesome and similar to the great complexifiers arguments. In this case, if a variable did not correlate with student outcomes, one could argue that the variable would have been significant if only the size of the variable or the number of dimensions or the unit of recording or something else had been different.

The above argument seems as unresolvable as the argument of the great complexifiers. The suggested additional procedures for encoding observations seems as plausible and researchable as the additional contexts suggested by the complexifiers.

The issue is further complicated by the also plausible idea that one type of unit of analysis unit (for example) may be most functional for one variable and another type of unit for another variable.

Some research seems called for to determine the functional value of some of the above questions. But, I don't believe that one can tackle or expect to tackle all these questions. The best one could hope for would be to focus on those issues in the above list which most people consider relevant, and such a needs assessment could be done by sending a checklist to a panel of experts.

II. Indexing implementation.

One could make a case that the indexing of program implementation is a fairly straightforward matter. One takes the behaviors considered important by the developer, develops an observation instrument, and uses the instrument to develop an index of implementation.

In practice, different investigators who are studying implementation are doing different things. Soar, for example, did not go to the program developers for lists of critical behaviors. Instead he chose observation instruments which he believed reflected the differences across eight Follow Through programs and used these results as an index of implementation. Soar first factor analyzed his results and then determined the range of classrooms within each program on the relevant factors. When the within program range was smaller than the across program range this was taken as indicating successful implementation. Using the Newman-Kehls procedure, Soar found a number of relevant dimensions on which programs differed and such differences usually reflected the a-priori orientations of the programs. Thus, implementation for Soar meant differentiation.

Stallings also used a differentiation procedure to index implementation. However, her observational instrument was constructed differently. Stallings first observed Follow Through classrooms and used her notes on the different models to construct her instrument. Following this, she asked each sponsor to select those variables considered most important for implementing their program and to further programs or control classrooms.

Siegel developed a set of implementation variables only for the DISTAR (or Oregon) Follow Through Model. Illustrative variables were:

teacher follows the program format when working with the entire group

correction procedure for mistakes when pupil does not understand teacher's signal

repeating task from beginning when pupil does not understand teacher's signal

ratio of attempts to obtain a unison response to the number of non-unison responses

Variables such as the above could be used to observe any program, but the behaviors are most likely to occur in the DISTAR program or in a similar structured, interactive program such as the Southwest Lab Communication Skills program.

These three investigators, all of whom were interested in indexing implementaion, have used three different procedures to do so. The variables each has selected differs from the others both in the range of events covered and the level of specificity. (Whether a greater range or a more detailed level of specificity is functional is an empirical question.) One might expect that other investigators would come up with still other procedures for indexing implementation. So how does one proceed?

12. Which problems should be studied and when?

The above question, raised by Joy during the meeting, seems very important. I would recommend that lists of research problems be developed and a panel, such as the one which met, attempt to see if they can assign priorities to the research problems.

Some research questions which I offer are:

a) in what settings should the research take place? (naturalistic, standard situation, specific curriculum product.)

b) what types of variables should be selected (those which focus on curriculum-emphasized activities, those which include general instructional variables).

c) what type of recording scale is most functional for what type of items? (frequency count by time, frequency count by natural unit, sign count, rating).

d) what variables are worth studying?

e) what contextual variables (e.g. pupil parent income, school location, curriculum, boy/girl ratio) are most important?

I am not sure how one would go about making decisions about priorities for studying the above issues. I recommend, for starters in a discussion, that (c), -- the recording scale -- and (d) variable selection be given top priority because solution of these issues is necessary for work on the next issues.

13. The need for a list of research issues.

The above issues and problems are certainly not exhaustive. I recommend that a list of issues and problems be developed and that a panel work on (a) defining the issues and (b) placing research priorities on these issues, and (c) suggesting research stragtegies.

14. A proposal for a data bank for secondary analyses.

If a data bank were available then many of the issues we raise could be subjected to empirical study. At the minimum, a data bank would include information on classroom transactions and on student outcomes. Transaction data could be on videotape or audiotape as well as in pupil questionnaires, observer ratings, and category counts. Bob Soar's material is an example of one type of material which could be included in such a bank.

A data bank need not be stored in any single place, but it is important that the items within a bank be assessible to researchers.

15. The importance of instructional research within curriculum programs.

Whether observational research within curriculum programs will yield better results than observational research which ignores program distinctions is a testable hypothesis. The argument is made here that some curriculum products provide teachers with tools which they would not receive if they worked without the program. To an unknown extent, these tools facilitate student learning.

It is thus hypothesized that instructional research which aimed at improving the impact of selected curriculum programs will be more effective than research which attempts to improve the general impact of teachers across programs. This is not to say that general instructional variables should not be studied (even within the context of specific programs), but, rather, that the payoff would be greatest when research takes place within specific programs.

A major reason for the above argument has been the research in Planned Variation Follow Through. At the summative level, certain types of programs have been consistently more successful in engineering pupil achievement than other programs. These findings suggest that these successful programs have been extremely successful tools for the teachers. Therefore, it is suggested that observational work designed to improve these tools would be a wise investment.